端奇

AI 应用开发实习生 / Python 后端实习生（AI 方向）

手机：13160051916　邮箱：qiduan332@gmail.com　GitHub：github.com/qd-maker　在线简历：cv.heyqi.xyz

南京工业职业技术大学 · 网络工程 · 2028届 · CET-6

求职意向

AI 应用开发实习生 / LLM 应用开发实习生 / Python 后端实习生（AI 方向）。大二网络工程本科，已完成两个可在线演示项目：保险条款 RAG 决策 Agent、多模型协作式 AI 对话平台。能围绕 MVP 完成需求拆解、API Contract、模型 API 接入、PostgreSQL/Supabase 落库、SSE 流式交互、Docker 部署与 README/测试记录沉淀。

教育背景

南京工业职业技术大学（本科）　　**网络工程**　　**2028届**

英语六级（CET-6），可阅读英文技术文档和官方 API 文档；相关课程：计算机网络、数据库原理、数据结构、Python 程序设计、Web 开发基础。

专业技能

- **后端与 API：**Python、FastAPI、Pydantic、REST API、SSE 流式接口、asyncio 并发任务、错误处理、限流、健康检查；能先写 API Contract 再落接口。
- **AI 应用工程：**有 RAG 项目实践，覆盖文档解析、Chunk、Embedding、pgvector 检索、Hybrid Search、RRF 融合、Rerank、引用溯源、结构化输出；有多模型编排、联网搜索注入、Prompt 拆分与共识生成实践。
- **数据与权限：**PostgreSQL、Supabase、pgvector、Supabase Auth、JWT、user_id 数据隔离、会话/消息持久化、基础 SQL 与迁移脚本。
- **前端与部署：**React、TypeScript、Next.js、Vite、Tailwind CSS、Zustand、EventSource 流式消费、Docker / Docker Compose、Git、pytest、项目 README 与部署文档。

项目经历

Insurance RAG Engine v2.0 / 保险产品决策 Agent　AI 应用开发 / 全栈开发　2026.01 - 2026.05

线上地址：insurance.heyqi.xyz | GitHub：github.com/qd-maker/insurance-rag

技术栈：Next.js 16、TypeScript、Supabase、PostgreSQL、pgvector、OpenAI API、Zod、SSE、LangFuse、RAGAS、Docker

项目简介：面向保险条款问答、产品解读、产品对比和风险审计场景，解决普通 ChatBot 在高可信场景下容易幻觉、跨产品串证据、结论无法回溯原文的问题；负责知识库入库、RAG 检索、Agent API、前端交互与部署文档。

- **4 模式 Agent API：**实现 `/api/agent/{ask,explain,compare,audit}`，使用 Zod 校验请求参数，统一输出结构化 JSON；关键结论绑定 `sourceClauseIds` 和条款片段，方便用户回看原文。
- **保险条款 RAG 链路：**完成 PDF 解析、语义分段、Overlap、产品别名归一和向量重建；检索阶段支持 pgvector Dense Search + 关键词召回，使用 RRF 融合候选结果，并按产品 metadata 过滤，减少跨产品污染。
- **产品对比与审计：**将两款产品按 6 个维度并行检索，再由单次 LLM 汇总推荐、风险提示和下一步行动；README 记录 compare 模式实测约 16-17s 完成。
- **工程化闭环：**提供 `/api/health` 聚合检查环境变量、Supabase、OpenAI、数据库、RAG Pipeline 和缓存状态；沉淀 Docker Compose、数据导入、缓存策略、日志分析、RAGAS/质量评估脚本与 GitHub Actions 评估流程。
- **产品取舍：**通过下拉选择约束产品范围，把 MVP 重点收敛到“已有产品回答可溯源、可比较、可审计”，提升演示稳定性，也减少无证据回答风险。

Quorum / 多模型协作式 AI 对话与共识生成平台　AI 全栈开发　2026.04 - 2026.05

线上地址：quorum.heyqi.xyz | GitHub：github.com/qd-maker/Quorum

技术栈：Python 3.12、FastAPI、Pydantic、React 18、TypeScript、Vite、Tailwind CSS、Supabase、PostgreSQL、SSE、OpenAI 兼容 API、pytest

项目简介：面向复杂问题讨论场景，搭建支持单模型聊天、多模型群聊讨论、联网搜索、图片/文本附件、追问、共识摘要和历史记录的 AI 应用；负责 FastAPI 后端、SSE 流式链路、多模型编排、用户隔离与演示模式。

- **接口与流式链路：**完成 `/chat`、`/discuss`、`/discuss/followup`、`/sessions` 等接口，Pydantic 校验请求模型；后端通过 `StreamingResponse` 推送 SSE chunk，前端用 EventSource 实时消费，并设置代理禁缓冲和心跳。
- **多模型并发编排：**4 个模型按 Round 1 / Round 2 并行讨论，支持自定义角色、图片输入和联网搜索上下文；使用 `asyncio.Queue` 汇总并发输出，保证前端按事件类型有序渲染。
- **共识生成优化：**把“全部发言一次性塞给共识模型”改为“各模型观点总结 + 共识合成”两阶段，将共识输入压缩到约 500-700 字；项目决策记录中统计 P95 时延由约 25s 降至约 6s。
- **真实故障处理：**针对第三方代理瞬时 429 / 502 / 503，加入 0.25s 错峰启动与 0.6s / 1.5s / 3.0s 握手重试；客户端断连时取消子任务，失败模型通过 `errors_summary` 显式透传。
- **多用户与测试：**基于 Supabase Auth / JWT / user_id 做会话隔离，支持 per-user API key 与模型映射、Demo Mode 和限流；编写 pytest 覆盖完整讨论、单模型报错兜底、搜索注入、追问流程和 Prompt 构建。